“DESIGNING EFFECTIVE MODELS FOR DOCUMENT IMAGE ANALYSIS BASED ON DEEP LEARNING”

Synopsis Submitted To

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY, NANDED, M.S. (INDIA)

For the Award of the Degree of

DOCTOR OF PHILOSOPHY

In Computer Science and Engineering under the Faculty of Science & Technology

Submitted by

Ms. Snehal P. Dongre

Under The Guidance Of

Dr. D. D. Doye
Professor,
Dept. of Electronics & Telecommunication
SGGSIET, Vishnupuri, Nanded
Nanded-431606, M.S.(India)

May 2022

Synopsis on

DESIGNING EFFECTIVE MODELS FOR DOCUMENT IMAGE ANALYSIS BASED ON DEEP LEARNING

1. Introduction and Motivation

In India, ancient manuscript recognition is drawing a lot of attention from researchers nowadays. Especially ancient manuscripts are rare and delicate records that also need huge space to store and preserve. To exploit the priceless information included in these documents, these documents are being digitized. Today most of the digital libraries publish the document over the web also. Due to increasing demand for digital libraries, document recognition and compression is becoming more popular in recent years. These documents are in the form of Printed and Handwritten Text. Current digital libraries use advanced recognition to extract the text. In Indian Language (IL) context most of the important documents are in printed or handwritten form. Therefore converting these documents into electronic form is a difficult task. Indian Language scripts are having different variations, hence there is high demand to build powerful recognition techniques to analyze the document contents and also require effective compression techniques to preserve and convert such documents to the editable form for most of the Indian language. Also the compression methods achieving high compression ratio is required for retrieval and accessing of these document images over the web. In Indian Language, documents with handwritten text have historical significance. Thus to preserve such documents for long run and for speedy communication over the web this data must be in compressed form.

Handwritten character recognition has always been a challenging task because of a variety of characters in different languages. It is the process of classifying handwritten characters into appropriate classes based on the features extracted from each character. In the past few decades, many feature extraction methods and classification algorithms were invented. The major role of text recognition and digitization was in document reading, signature identification, postal code verification, handwritten notes digitization. To support and improve performance of these applications, recognition capability has been expanded in many other dimensions like mode of writing, structure of scripts, types of documents, etc. Every script in the world has a different structure; therefore a specialized technique is engaged to recognize the respective script. Some of the scripts can have similar script writing style, so, recognition technique to these scripts can be somehow similar.

Handwritten text recognition is more complicated than machine-printed script due to variations in writing by different users. The character set of Indian languages is large and comprises more complex characters when compared to the European scripts. Indian scripts are slightly multifarious with huge character sets, complex character shapes, variation in writing styles, etc making it difficult to achieve a good recognition accuracy rate. Many attempts were made to gain good accuracy in handwritten text recognition for Indian languages like, Pali, Tamil and Telugu and many more. Due to complex orthography present in the scripts, recognizing Marathi and Pali characters is an interesting yet challenging task. The complex architecture of characters is another major reason that makes the handwritten character recognition task more complicated. Research in this direction focuses on segmentation procedures, feature extraction procedures, and classification algorithms. Character recognition itself is a challenging problem because of the variation in the font and size of the characters, type of documents, writing style, etc. The feature extraction method and the classifier are commonly used for extracting the features as well as for the recognition of Marathi and Pali language characters in which it successfully segments out the lines and words of the Marathi and Pali documents. OCR is a very convenient way to digitize printed or handwritten documents so that they can be readily available for editing and formatting. To store and transmit such handwritten documents ultimately needs more storage space in the hard drive of the computer. Compressing such documents can solve this issue to a great extent.

Compression is the process of converting a data set into a code to save the need for storage and transmission of data. Data compression is widely used by the community because through a compression we can save storage space. Data compression can also speed up a transmission of data from one place to another. In performing a compression requires a method of data compression that can be used. Data that can be compressed not only text data but images. With the compression technique one can achieve efficiency in saving time and storage that exist in memory storage. Each compression algorithm has its own features in terms of efficiency, response time and Compression Ratio (CR). By implementing and comparing different recognition techniques and measuring CR in each case, the efficiency of applied compression techniques can be compared. Data compression technique is divided into two type’s namely lossy compression and lossless compression. Algorithms such as Huffman, Shannon Fano, Tunstall, Lempel Ziv Welch and Run-Length Encoding are lossless compression algorithms. Each method has the ability to perform a different compression. A compression will be done to facilitate the transmission of a file with a large size and contains many characters. The workings of compression techniques are based on patterns of repetition in the data and replace it with

certain symbols.

Decompression results depend on the nature of the compression used, namely Lossless Compression or Lossy Compression. If a lossless compression technique has been performed on a text, the original text can be recovered exactly from the decompressed file. After decompressing such handwritten documents to the human editable format, they must produce high quality text documents and also should have the ability to produce high accuracy results. To design and develop such a technique is really a challenging task in the field of document digitization.

There are various techniques available for text recognition. In the proposed work, the ancient handwritten manuscripts are recognized by using proposed recognition techniques for a particular category and by applying an efficient compression algorithm; the document can be preserved with less storage space, high compression ratio and high response time. Also the result of using earlier techniques can be compared for performance analysis. The work will also serve to enhance the collection, distribution, analysis, and overall scientific worth of historic data.

2. Literature Survey

The authors in [1] devised a framework that can recognize similar handwritten digit strings. The main purpose of their research was to construct an automatic digit recognition system for text pictures using a machine learning approach. They focused on a proposed objective to use a recognition system to identify distorted handwritten digits, where the sub images of attached digits, disjoint digits and overlapped digits were developed. Then using sub images, a classifier was trained. The classifier was then used to classify individual pixels in a picture. Because of divided numbers, conflicting numbers, identical numbers, digit ambiguity, undefined string duration and over and under segmentation, it was a big hurdle to segment adjacent numbers. They used water reservoir segmentation of two linked handwritten digits using SVM ANN and CNN methods to decide the proper methodology for the isolation of two distinct digits. Further in their paper they analyzed the performance of a system using SVM, ANN and CNN with respect to precision and efficiency in text picture recognition.

Authors in article [2]Ashay Singh and Ankur Singh bist in their paper presented a comparative analysis of recent techniques for character recognition with the clear purpose of to identify the impact of machine learning in the domain of character identification. They put up their views that traditional machine learning techniques like a neural network, support vector machine, random forest, etc. have been used as classification techniques. They extended their work in the progress of the advancement in the field of computer hardware and efficient research in the artificial

intelligence field that leads in emerging to deep learning algorithms. They implemented a system using deep learning for character identification. They also depict how various functions improve the performance in the field of pattern recognition over time. The primary purpose of their paperwork is to encourage young researchers towards the domain of DL and thus learn and work towards achieving novelty work in the field.

Authors in the paper [3] proposed a recognition system to process input image, extraction of features, classification scheme and training of system to acknowledge the text. During the approach, they trained the system to seek out the similarities, and also the differences among various handwritten samples. The application takes the image of a hand transcription and converts it into a digital text. They focused on classification of characters. They implemented through the conventional neural network. The accuracy obtained in the experiment was above 90.3%. The proposed algorithm provide with both the efficiency and effective result for the recognition. The system gave best accuracy for the text which has less noise. The accuracy completely depending on the dataset if we increase the data, we can get more accuracy.

The authors in their survey paper [8] reviewed deep learning-based methods, datasets, and metrics for six main document image enhancement tasks, including binarization, debluring, denoising, defading, watermark removal, and shadow removal. They presented the details of the recent works for each task and discuss their features, challenges, and limitations. They introduced multiple document image enhancement tasks that have received little to no attention, including over and under exposure correction, super resolution, and bleed-through removal.

The authors of journal article [5] provided a pragmatic definition of historical documents from the point of view of the research in the area that was addressed in the research. Further they described the different input-output relations that are expected from the used deep learning approaches and also accordingly described the most used models. The authors discussed the research datasets published in the field and their applications. In their paper they addressed specific problems in the overall process of Document Image Analysis(DIA) like Building Collections in Digital Libraries, Writer Identification, Manuscript Dating, Estimation of Transcription Costs, Pre-Processing, Image Enhancement, Image Binarization, Layout Analysis, Table Detection/Recognition, Text Line Segmentation, Baseline Detection, Character and Symbol Recognition, Keyword Spotting, Text Detection, Character Recognition. They also focused on different AI, DL CNN architecture that can be used for DIA. The analysis proved that the latest research is a leap forward since it is not the simple use of recently proposed algorithms to previous problems, but novel tasks and novel applications of state of the art methods that are to be considered. At the end of the paper they explored some potential future trends that can represent a stimulus for innovative research directions in the field of DIA.

The authors of the article [4] deduced a framework proposing character recognition system

for documents written in Kannada language. They proposed the technical details of design and implementation of the CNN model for the implementation and used the Chars74K dataset used for training the model. They used a non- non-local means algorithm which deals with Gaussian noise which arises during scanning the document. Non-local means filtering calculates the target pixel to estimate the original image by computing the mean from the surrounding pixel. The built model has achieved accuracy of 98% for the document containing non-overlapping lines of characters.

Authors of the article [7] proposed a model for segmentation document image into text-lines and words in JPEG compressed domain. They attempted their work to segment a JPEG compressed printed text document image into text lines and words without fully decompressing the image With JPEG compression, the non-overlapping 8*8 DCT blocks encode text contents of two adjacent text-lines and words without leaving any visible clue for segmentation. The paper proposed two stage algorithms for segmentation of text-lines and words by intelligently analyzing approximate text-line and word boundaries using the DC coefficient during the first stage. In the second stage, AC coefficients of selected DCT blocks are used to extract exact line and word boundaries.

The research scholars in the article [6] proposed a devised method, compression-based di-rect processing, to enable high performance document an-alytics on compressed data. By enabling efficient direct processing on compressed data, the method saves 90.8% storage space and 77.5% memory usage, while on average, speeding up the analytics by 1.6X on sequential systems, and 2.2X on distributed clusters. The paper presents how the method can be materialized on Sequitur, a compression method that produces hierarchical grammar-like representations.

The article of [9] introduces Four Side Border Features (FSBF) technique which extracts features of pre-processed images. Further, these features are used to find unattended information of the image. This information is processed by the Segmented Binary Block (SBB) method. Some characters in Kannada script are more complicated in their structure; therefore, Segmented Four Side Border Features (SFSBF) method is required for feature extraction. All these features are recognized by the Support Vector Machine (SVM) with 3-fold cross validation technique. The generated results out of our experiment were compared with similar kinds of algorithms which show that accuracy of character recognition is 92%.

3. Objectives of Research/Proposed Hypothesis

There is lots of research work available on Document Image Analysis. Generally a complete document may contain images, logo, text, numbers, and signature either present in printed or handwritten form. Most of the work generally focused on specific contents of the document either text, character or numbers. The research proposed here is aimed towards designing an effective deep learning model for Document Image Analysis with significant recognition rate and efficiency.

Objectives are as follows:

(a) To Study various challenges in available research work on Document Image Analysis techniques.

(b) To Design a model to recognize complete documents irrespective of its content by making use of DL technique.

(c) To compare performance analysis with the compressed and uncompressed document image. .

4. Methodology to be adopted:

The primary objective of this research is to develop a system to recognize the document irrespective of its content. Initially literature review is to be carried out to gain information about machine learning data sets and familiarize with the various machine learning techniques available to train the data set. It can be conducted to get familiarized with the multiple databases used for research and examination. In the next stage, certain work will be carried out for understanding various data pre-processing, segmentation, and machine learning methods to be used in the analysis. There are certain methods to be studied for the implementation of the proposed model, like Handwritten Document Segmentation, Handwritten Document Classification, Machine Learning Methods, Deep Learning, and Image Analysis on Text Files, Support Vector Machine, Artificial Neural Networks, Conventional Neural Networks and Preprocessing Document. Training model is to be built using deep learning techniques. The ultimate aim of the proposed work is to build an effective automated model for complete document analysis using machine learning methods to recognize overlapping handwritten numbers, digit strings and dates. The proposed research work would examine the adoption of emerging machine learning and deep learning approaches to various datasets, linke SVM, ANN and CNN and their performance would be analyzed to drive efficiency tests. The model is then compared with the compressed document analysis with respect to recognition rate and time and space required for recognition. With recent advancement in Deep Learning, methods are proposed to enhance the quality of handwritten document images. The proposed research

will review deep learning-based methods, datasets and metrics for certain document image enhancement tasks including binarization, deblurring, denoising, defading, watermark removal and shadow removal.

5. Importance of study/ Society Application:

1. International Status:

The proposed document recognition model can be used on an international level as document analysis has become a popular subject as it may cover ancient documents to be preserved in its original form. Also preserving such a document is very time consuming and again it will require more space for its storage. The proposed model is helpful in this regard as it is to be developed using DL to provide a high recognition rate with comparatively less time and space requirement.

2. National Status:

The proposed work is also helpful in preserving historical handwritten documents and also storing it with less space. The model can be used commercially to recognize the complete document irrespective of its content. The work will also serve to enhance the collection, distribution, analysis, and overall scientific worth of historic data.

3. Significance of the Study:

Certainly the proposed model would have its own significance in the field of Document Image Analysis as the proposed system can recognize the contents of documents with less time and with more accuracy using DL as DL is having its own advantages over existing systems built in the area of DIA. Nowadays any document can be scanned and processed to recognize and find its usability.

6. Proposed Work Plan / Formulation and Structure of Study Work:

Phase	**Work To Be Carried Out**
I	- Literature Survey, Course Work - Formulation of Problem Definition, Objectives and Feasibility Study.

II	- Survey paper publication - Design and development of DIA recognition model using ML and DL
III	- Analyze performance evaluation. - Publications in reputed Journals and Thesis Writing

7. References:

[1] Arti Gupta, Rohit Miri and Hiral Raja, "Recognition of Automated Hand-Written Digits on Document Images Making Use of Machine Learning Technique", European Journal of Engineering and Technology Research, ISSN:2736-576X, Vol-6, Issue-4,May 2021.

[2] Ashay Singh , Ankur Singh Bist "A Wide Scale Survey on Handwritten Character Recognition using Machine Learning" 2019, IJCSE Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

[3] Sri. Yugandhar Manchala, Jayaram Kinthali, Kowshik Kotha,Kanithi Santosh Kumar, Jagilinki Jayalaxmi"Handwritten Text Recognition using Deep Learning with TensorFlow", International Journal of Engineering Research & Technology (IJERT) Vol. 9 Issue 05, May-2020

[4] Askok K, Krishnappa H K, "Kannada Handwritten Document Recognition using Convolutional Neural Network", 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018, ISBN:978-5386-6078-2, IEEE Proceedings.

[5]Francesco Lombardi, Simone Marinai, "Deep Learning for Historical Document Analysis and Recognition—A Survey", J. Imaging 2020, 6, 110; doi:10.3390/jimaging6100110 www.mdpi.com/journal/jimaging

[6]Mohammad Hosseini "A Survey of Data Compression Algorithms and their Applications",https://www.researchgate.net/publication/270408593

[7] Bulla Rajesh, Mohammed Javed, P. Nagabhushan, Watanabe Osamu,Segmentation of Text-Lines and Words from JPEG Compressed Printed Text Documents Using DCT Coefficients, 2375-0359/20/$31.00 ©2020 IEEE DOI 10.1109/DCC47342.2020.00083

[8] Zahra Anvari, Vassilis Athitsos, "A SURVEY ON DEEP LEARNING BASED DOCUMENT IMAGE ENHANCEMENT", A PREPRINT - JANUARY 4, 2022

[9] Dr. Balaji Rajendra Bombade, Dr. D.D Doye "Handwritten Kannada Character Recognition with Border Features" Journal of Applied Science and Computations Volume V, Issue XII, December/2018

[10]Feng Zhang, Jidong Zhai, Xipeng Shen, Onur Mutlu, Wenguang Chen, "Efficient Document Analytics on Compressed Data: Method, Challenges, Algorithm, Insights", Proceeding of the VLDB Endowment, Vol-11, No-11 2018.

[11]Luluk Anjar Fitriya, Tito Waluyo Purboyo, Anggunmeka Luhur Prasasti "A Review of Data Compression Techniques"International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 19 2017 pp. 8956-8963 © Research India Publications. http://www.ripublication.com

[12] Milind Bhalerao, Sanjiv Bonde and Madhav Vaidya “Frequently Used DevanagariWords in Marathi and Pali LanguageDocuments” Springer Nature Singapore Pte Ltd. 2019

[13] Dr. Suresh Limkar, Gautami Mudaliar, Sneha Kulkarni, Neha Rathod,Tejasvi Gadakh, Sanaya Shah Optical Character Recognition for Marathi Language using Deep Convolutional Neural Network” Journal of University of Shanghai for Science and Technology, D.O.I - 10.51201/12488 https://doi.org/10.51201/12488, Volume 22, Issue 12, December – 2020

[14] Zi-Rui Wang, JunDu “Joint architecture and knowledge distillation in CNN for Chinese text recognition” 2020 Elsevier Publicatios, Pattern Recognition, Volume 111, March 2021, 107722

[15] Yu Zhou, Shuang Liu,Yongzheng Zhang,Yipeng Wang,Weiyao Lin,” Perspective Scene Text Recognition with Feature Compression and Ranking”, Asian Conference on Computer Vision ACCV 2014: Computer Vision - ACCV 2014 Workshops

[16] Rameshwar S. Mohite, Balaji R. Bombade ” Challenging Issues in Devanagari Script Recognition” IJCTA May-June 2014

[17] R. Jayadevan, S. R. Kolhe, P.M. Patil and U. Pal, "Offline Recognition of Devanagari Script: A Survey", IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, 2011.

[18] Kamble, P.M., Hegadi, R.S.” Handwritten Marathi character recognition using R-HOG Feature”. Procedia Comput. Sci. 45, 266–274 2015

[19] G. G. Rajput, Rajeswari Horakeri, “Shape Descriptors based Handwritten Character Recognition Engine with Application to Kannada Characters”, International Conference on Computer & Communication Technology (ICCCT), pp 135-141, 2011.

[20] Umapada Pal, RamachandranJayadevan, Nabin Sharma, Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques, ACM Transactions on Asian Language Information Processing (TALIP), Volume 11 Issue 1, article 1,pp 1-35, March 2012.

[21] SandhyaN, Krishnan R, and D.R.Ramesh Babu, Feature Based Kannada Character Classification Method of Kannada Character Recognition, International Journal of Scientific & Engineering Research, Volume 5, Issue 2, pp 729-733,February-2014.

Ms. Snehal P. Dongre
***R**esearch **S**cholar*
Dept. of Computer Science & Engineering
SGGSIET, Nanded

Dr. D. D. DOYE
RESEARCH SUPERVISOR
Professor, **D**ept. of
Electronics & **T**elecommunication **E**ngineering
SGGSIET, Vishnupuri, Nanded,
NANDED – 431606 MS (INDIA)